When Style Transfer Meets Retrieval

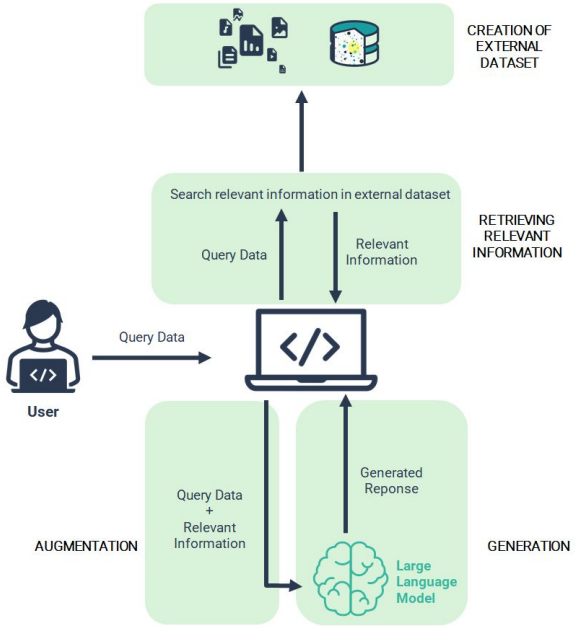
Presented by
Chen-En Ma, Jiayi Shen, Hongwei Liao

Introduction

Retrieval-Augmented Style Transfer

A **novel framework** that combines **retrieval-based augmentation** with **diffusion models** to achieve efficient and flexible **style transfer**.

Retrieval-Augmented Generation



Style Transfer

Base Image

+

Style Image

Transferred Image =

Problem

The success of style transfer relies heavily on the availability of appropriate style images, **which may not always be readily accessible.**

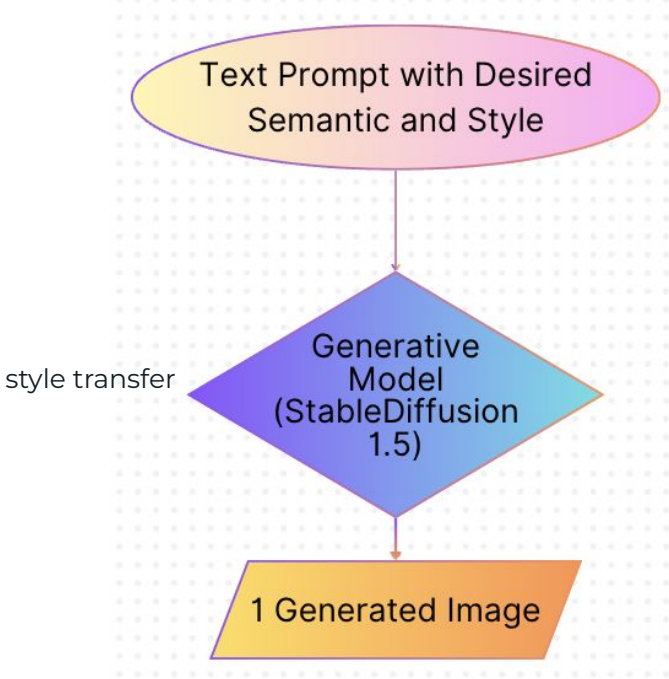
How we solve it:

1. Trained a representation model to foster text-to-image search for style images.
2. Leveraged natural language prompts to retrieve relevant style images.
3. Fed retrieved style images to diffusion models to complete stye transfer.

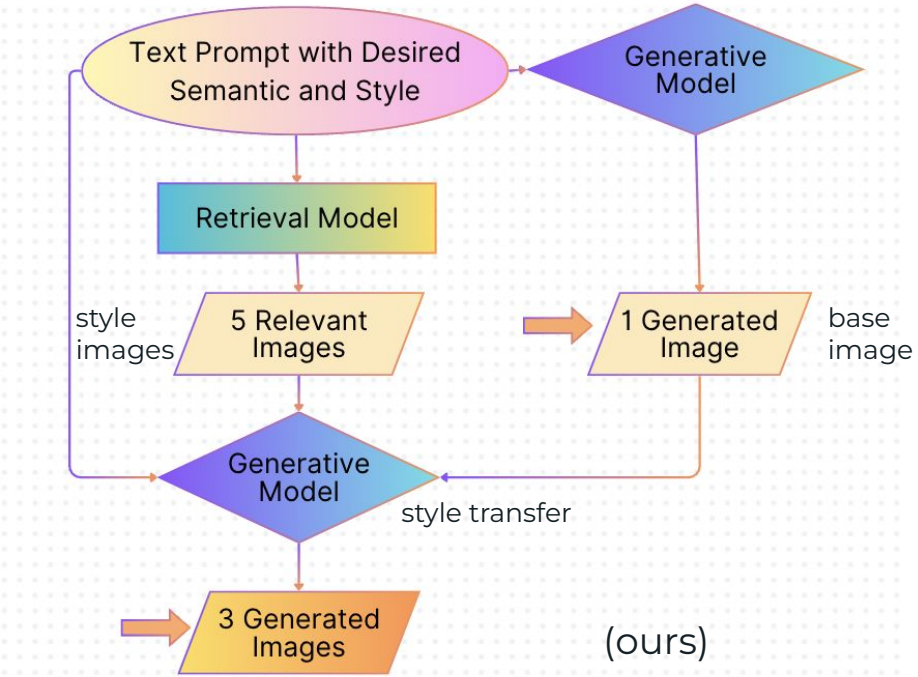
Methodology

Methodology Overview

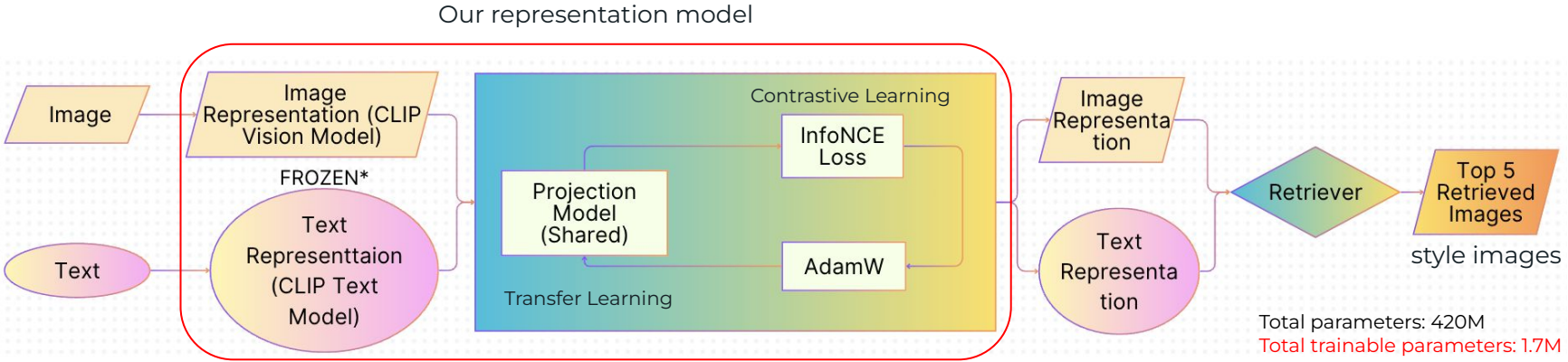
No Retrieval



Retrieval + Style Transfer



Text-to-Image Retrieval



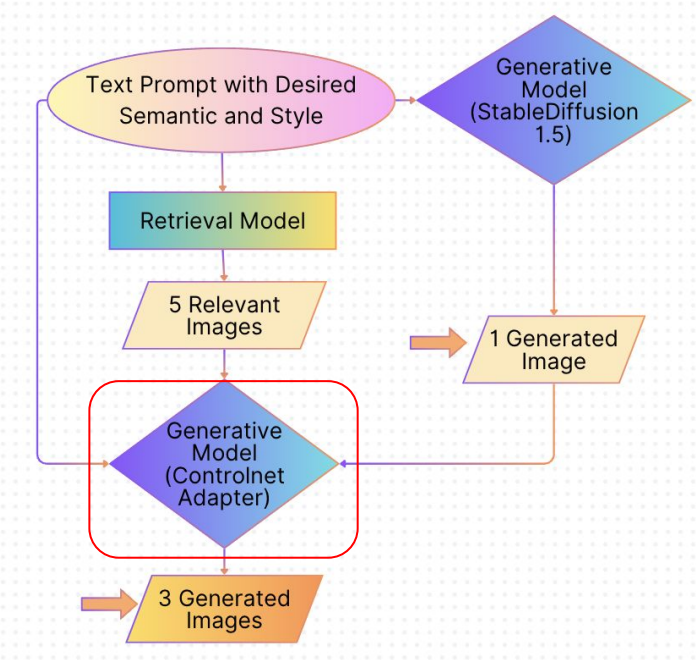
Retriever: similarity-based search and reranking function

Model details

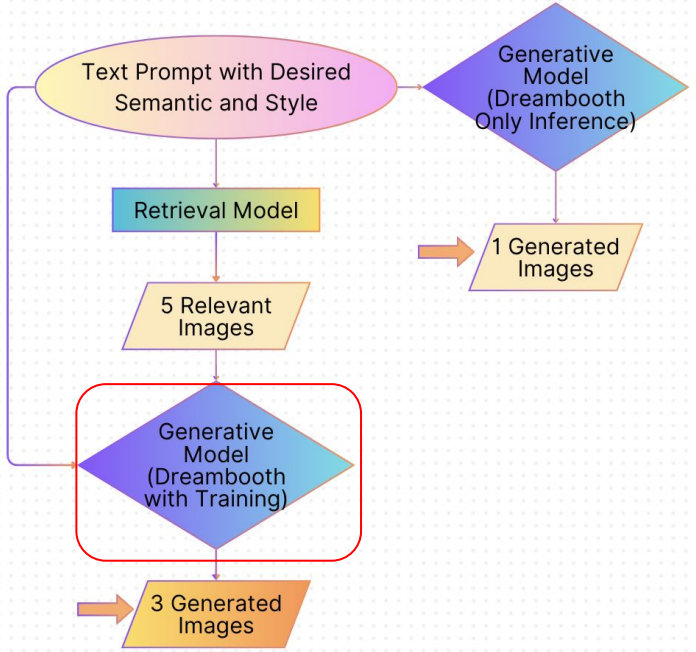
Tokenizer/Processor:	CLIP Processor
CLIP Variant:	Long CLIP (“zer0int/LongCLIP-GmP-ViT-L-14”)
Projection Model:	3 * Fully-Connected Layers Hidden Dimension = 768 Dropout Rate = 0.3
InfoNCE Loss:	Temperature = 0.1 Negative Pair Number = 5
AdamW:	Learning Rate = 1e-4
Retriever:	Sentence Transformers’ Semantic Search K = 5

Retrieval + Style Transfer

Retrieval + ControlNet



Retrieval + DreamBooth



Our Dataset

SemArt: 21000’ish fine art samples with descriptions

Title: Still-Life
Author: Willem van Aelst
Type: Still-Life
School: Dutch
Timeframe: 1651-1700

The painting depicts a still-life with roses, tulips and other flowers resting on a ledge. It demonstrates the elegance, refinement, and technical brilliance cultivated during the painter's formative years in Italy.

Results

Retrieval Outputs

prompt:

Create an artwork with Northern Renaissance-style inspired by the compositions of Pieter Bruegel the Elder, reinterpreted by Baltens. Emphasi

ground truth image index: 9
retrieved image index: [1739, 9, 1615, 1493, 1498]
ground truth image:

retrieved images:





Retrieval Metrics

	Accuracy @ 5 ↑	MRR ↑
CLIP Zero-Shot	0.57	0.22
CLIP w/ Transfer Learning	**0.91**	**0.81**

The higher, The better.
Accuracy@K: 1 if the ground truth context is retrieved in top K relevant context else 0
MRR: avg position of the ground truth image; 1/(the rank of ground truth image)

5 retrieved images for style transfer

Prompt:
Create an artwork of a **joyful trip** with Northern **Renaissance-style** inspired by the compositions of Pieter **Bruegel the Elder**, reinterpreted by **Baltens**. Emphasize intricate detail, narrative complexity, and vibrant scenes of everyday life. Incorporate **dynamic groupings** and a rich, textured atmosphere, reflecting the influence of Bruegel's 1564 painting while showcasing Baltens' distinct artistic vision and interpretation.

	Adapter	Dreambooth
text only		
text + retrieved images		

5 retrieved images for style transfer

Prompt:
Create an artwork of **company meeting** with people wearing **black suits**, using **Renaissance-style** depiction inspired by Leonardo da Vinci's works. Emphasize **central perspective and symmetry**. Incorporate dynamic compositions, **vibrant colors**, and symbolic details that reflect Leonardo’s innovative interpretation of the Gospel narrative.

	Adapter	Dreambooth
text only		
text + retrieved images		

5 retrieved images for style transfer

Prompt:
Create an artwork of **train carriage with Dutch Golden Age-inspired scene** in the style of Pieter de Hooch, blending still-life qualities with precise perspective, and luminous colors. Capture a serene, frozen moment imbued with timelessness, where the interplay of light and structure evokes a **sense of quiet eternity**.

	Adapter	Dreambooth
text only		
text + retrieved images		

CLIP Img2Img Similarity Score

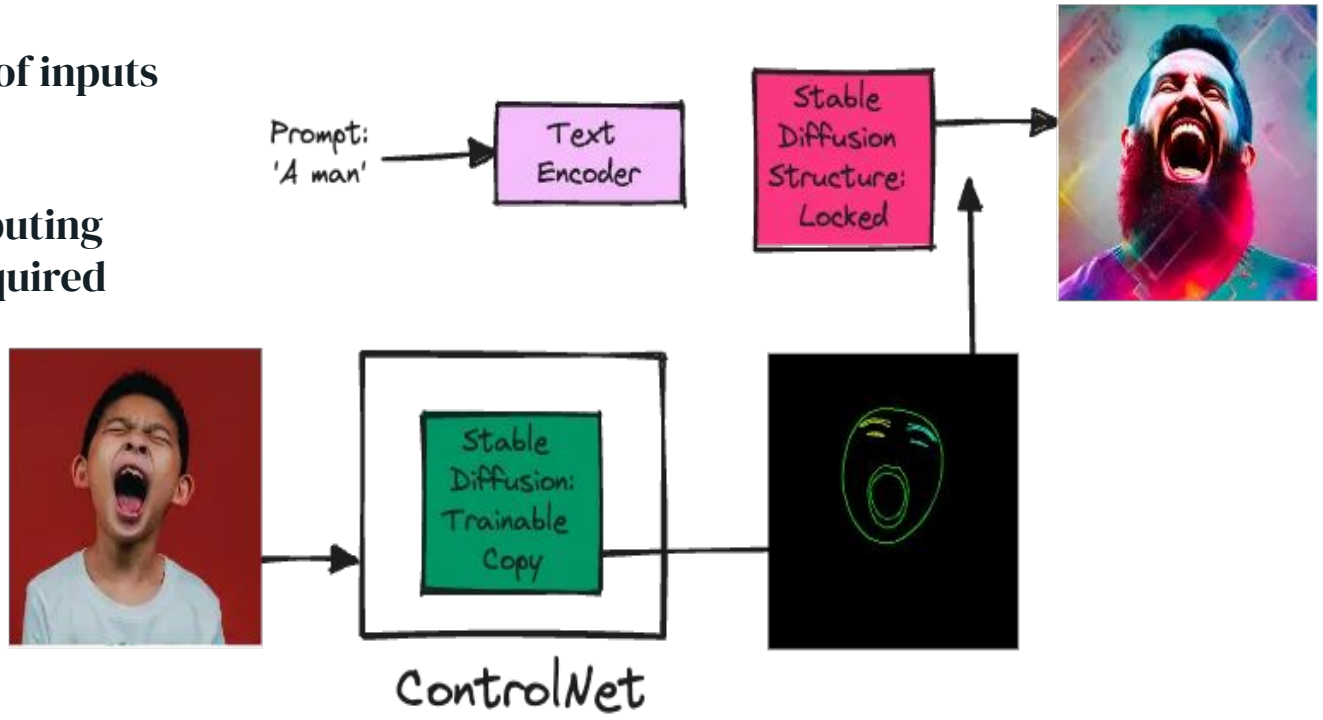
	Adapter	Dreambooth
Text Only	0.669	0.542
Text + 5 Retrieved Image	**0.728**	0.533

- Generated image vs Retrieved image
- Calculate the mean of pairwise cosine similarity of the CLIP image embeddings

Lessons Learned

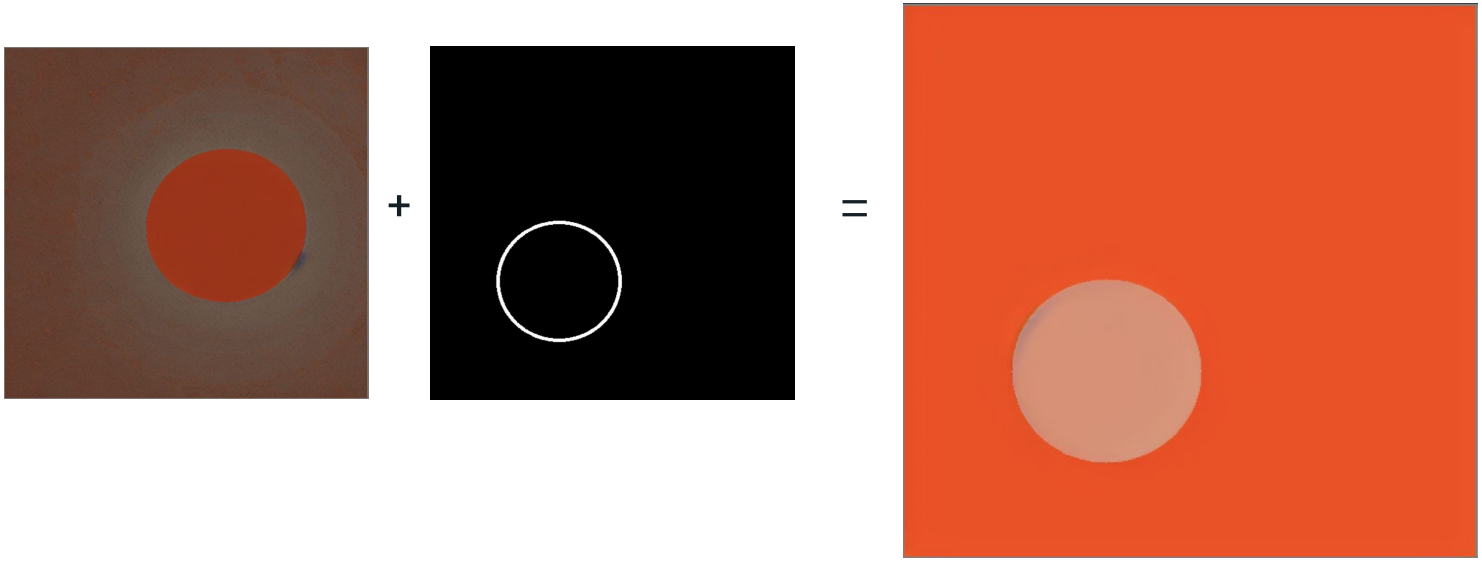
~~Training a Controlnet~~

- Limited number of inputs and outputs
- Substantial computing resources are required



~~Training a Controlnet~~

Text: Dark Salmon Circle with orange red Background



Therefore, training a ControlNet is impractical for this project

Future Work

- Train our own style transfer model.
- Continue fine tuning our retrieval model's to create better representations.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2020.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[3] N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," 2018.

[4] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[5] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2022.

Thanks!

Any questions?